MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2017/2018

TMA 7021 – DATA MINING AND ANALYTICS

(All sections / Groups)

1 FEBRUARY 2018
8.00 p.m – 10.00 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of **5** pages with FOUR questions only.
2. Answer **ALL** questions.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) CART is categorized as an *eager learner* while Naïve Bayes Classifier is a *lazy learner*. Explain the characteristics of these learners.

[3 marks]

(b) Outlier analysis is important in identifying abnormal patterns. Define outlier and name THREE(3) different approaches for outlier analysis.

[3 marks]

(c) A local bank is launching *cross-selling* campaign. Cross-selling is the practice of selling additional products to an existing customer. That is, *a customer who purchase product A and B will likely to purchase product C*. What is the most appropriate unsupervised machine learning technique that uses transactional data as input?

[3 marks]

(d) Name any TWO(2) programming languages or software tools that are suitable for data mining and analytics.

[1 mark]

Continued...

QUESTION 2

Study the following table carefully and answer the questions below:

Table name = **dt_Marks**

Name	Age	CGPA	Marks
Ting	20	3.5	95
David	20	3.5	90
Chong	25	3.2	89
Richard	18	3.1	88
Abdul	18	3.1	98
YeeLing	22	2.6	88

(a) Calculate the similarity for (Chong and YeeLing) and (Chong and Abdul) using *Euclidean* distance measurement method. What can you conclude?

[5 marks]

(b) Assuming that you are requested to perform a mark prediction for a new student given information about age and CGPA, what would be the most appropriate machine learning method for this purpose? Explain your choice.

[4 marks]

(c) The dataset above can be clustered into a few clusters. However, determining the number of clusters is difficult for *k-means* algorithm. What method can you use to determine the number?

[1 marks]

Continued...

QUESTION 3

(a) The table below shows the point of interest given a particular area.

Location ID	Business Type
1	Laundry, clinic, fruit shop
2	Fruit shop, clinic
3	Restaurant, school
4	Laundry, clinic, fruit shop
5	Clinic, laundry

(i) Calculate the *support* for the following item-set:

`{restaurant, school}`

[2 marks]

(ii) Calculate the *confidence* for the following rule:

`{fruit shop, clinic}` → `{laundry}`

[2 marks]

(iii) Calculate the *Lift* for the following association rules:

`{laundry, clinic}` → `{fruit shop}`

[2 marks]

(b) From your findings obtained in (a), assuming that you observed a fruit shop and a clinic in a row of shops, could you open a laundry business next to them? Explain your answer.

[2 marks]

(c) Define *Lift* in Association Rule Mining. What can you conclude if Rule1 has a lift of 3.00 while Rule2 is 1.50?

[2 marks]

Continued...

QUESTION 4

(a) You are requested to design and draw a Bayesian Network to help recommending the *decision* for a bank in provision of commercial loan. The nodes *Age*, *Company Type*, *Personal Loan*, *Business Type*, *Account Balance*, and *Payment on Schedule* are evidential nodes. The design of Bayesian Network has to support both diagnostic and predictive reasoning.

[4 marks]

Age	Company Type	Personal Loan	Business Type	Account Balance	Payment on schedule	Decision
5	IT	No	Local	10000	Yes	Accept
6	F&B	Yes	International	2000	No	Reject
10	Service	Yes	local	100000	Yes	Accept
...	...	...	...	...	...	...

(b) Discuss FOUR(4) characteristics of a domain where Bayesian Networks can be deployed.

[2 marks]

(c) In text mining, *Term Frequency-Inverse Document Frequency* (TF-IDF) is often employed for information retrieval. Define TF-IDF and how to interpret the results from TF-IDF.

[4 marks]

End of pages.